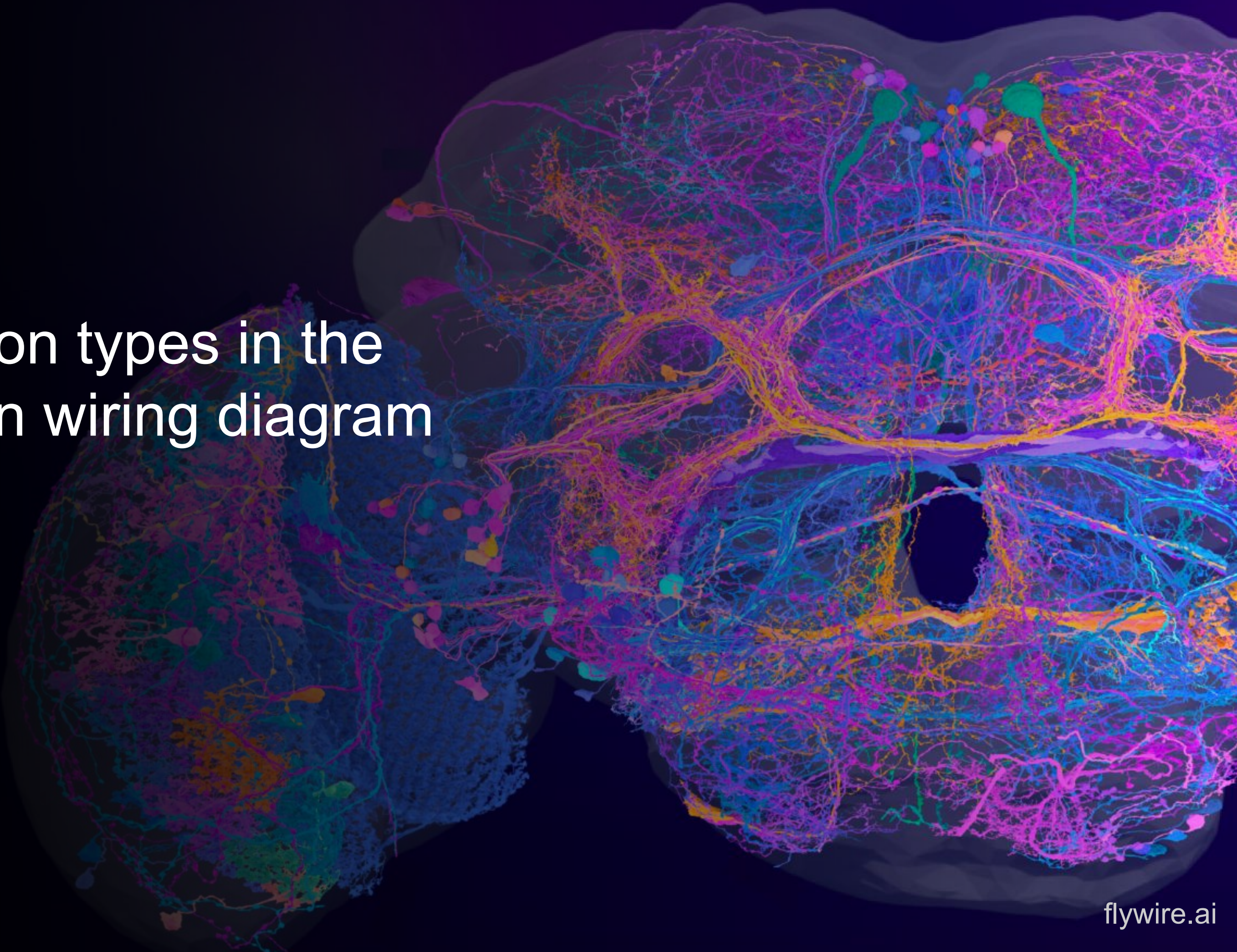
Predicting neuron types in the *Drosophila* brain wiring diagram

Ernesto Bocini
Sibo Wang



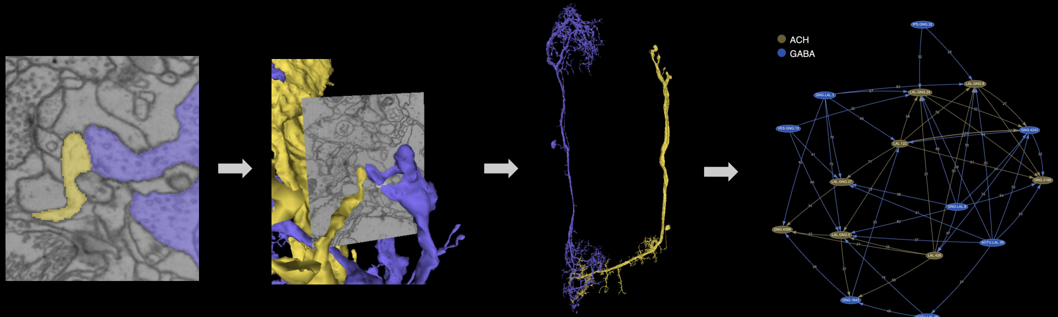
flywire.ai

FlyWire: The *Drosophila* brain connectome

Synapse-level reconstruction of *all* connections in the *Drosophila* brain

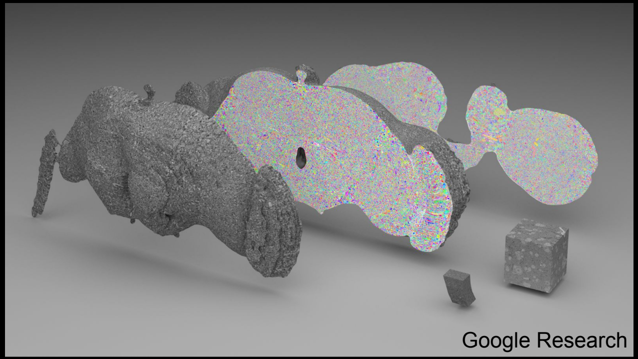
Derived from nm-scale electron microscopy data

Drosophila melanogaster



2D processing → 3D segmentation → Whole-neuron reconstruction → Graph representation

The complete wiring diagram of the whole brain

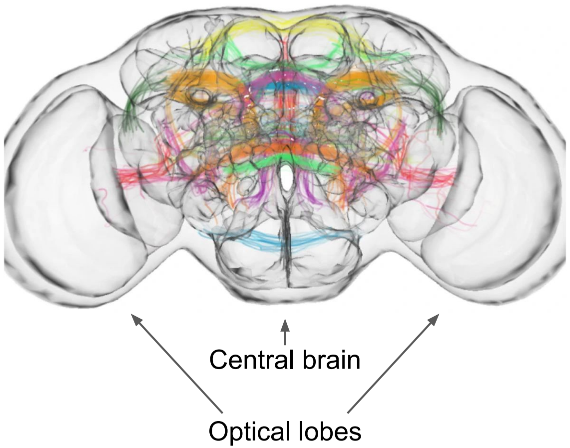


Dorkenwald et al. FlyWire: online community for whole-brain connectomics. *Nat Methods 19*, 119–128 (2022).

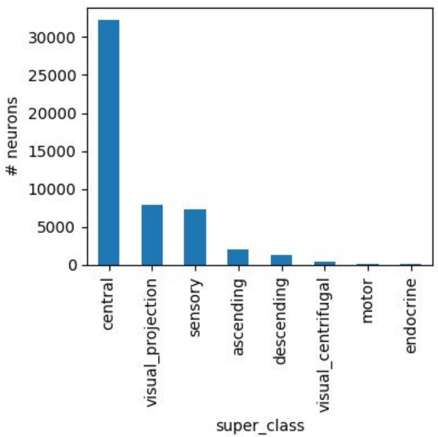
Nodes

- Each node represents a neuron (51,405)
- We excluded neurons in the optical lobes

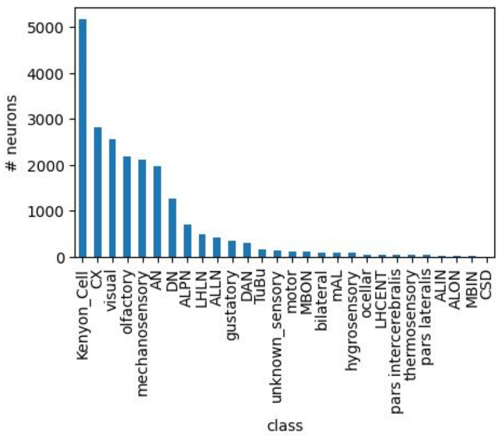
Figure: Chuang 2015, *BMC Neurosci*



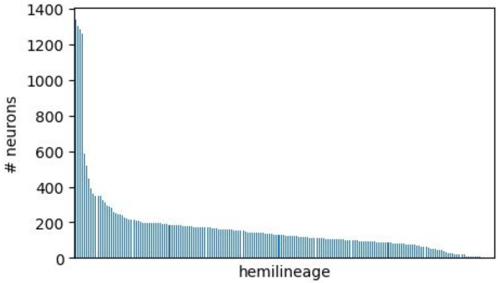
Neuron superclasses



Neuron classes



Hemilineages



Edges

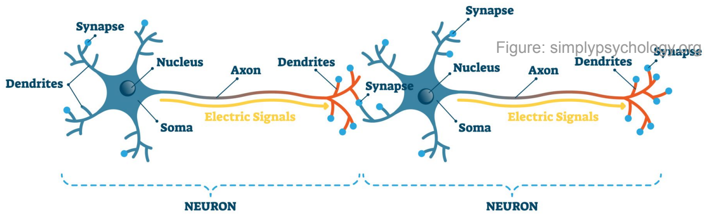
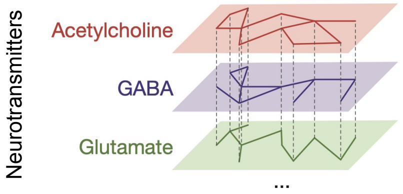
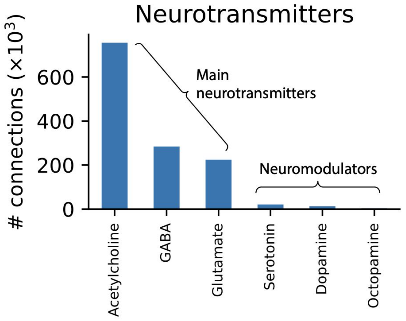


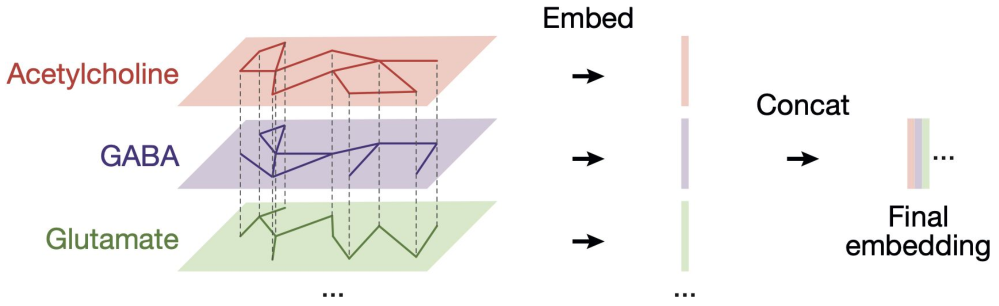
Figure: simplypsychology.org

- We consider the interaction among neurons via chemical synapses.
  - 3 main neurotransmitters: Acetylcholine, GABA, glutamate
  - 3 neuromodulators: serotonin, dopamine, octopamine
- This can be modeled as a multi-layer graph
  - **Biological scenario:** Neuron A synapses onto neuron B with N synapses using neurotransmitter type T.
  - **Graph formulation:** A directed edge from A to B with weight N on layer T.
- Total number of edges = 1,301,936



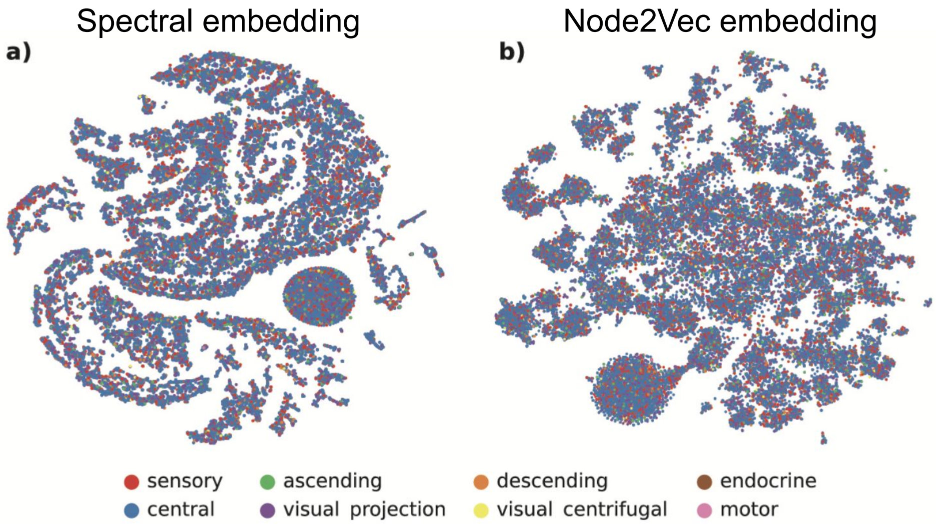


Building the node embedding



- Independently build node embeddings for each layer (neurotransmitter)
  - Spectral embedding using the directed Laplacian (Chung 2005)
  - Node2Vec (Grover & Leskovec 2016)
- Concatenate embedded vectors together for all layers

Building the node embedding



Visualization: clustering of nodes based on connectivity

Observation: naive unsupervised clustering does not segregate node classes

Further exploration: Analysis on the distribution

Power Law distribution: common in real world graphs.

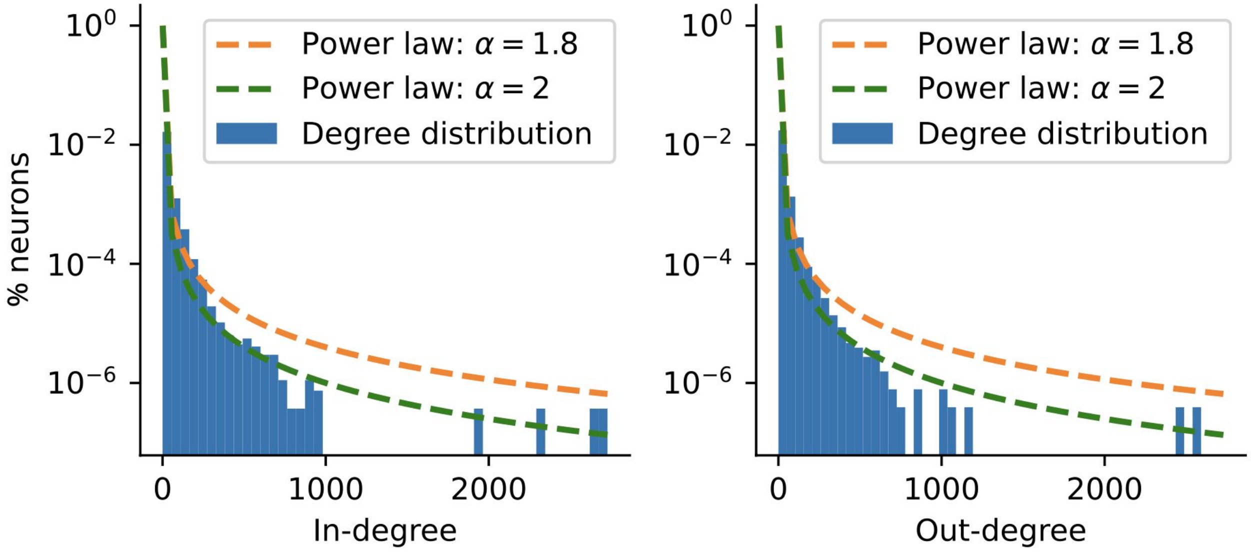
First moments:

- 22.11
- 22.10

Second moments:

- ~ 2600
- ~ 2000

α in the scale-free regime

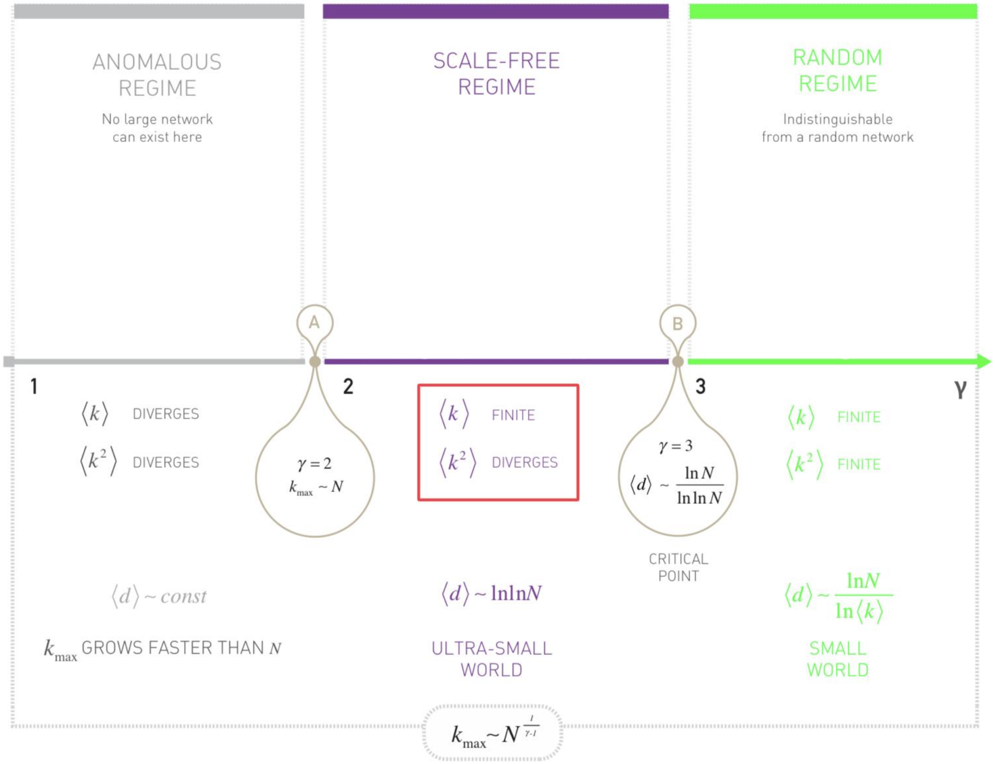


Other properties of the Graph

- Density = M/[N(N-1)] = 4.5 x 10 $^{-4}$
- Average Clustering Coefficient = 0.144
- Average distance ≃ 4.21

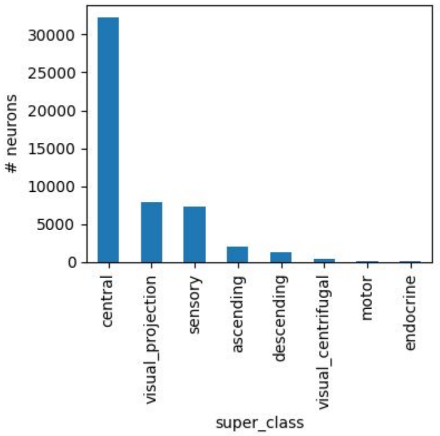
log(N)/log⟨k⟩ = 3.50

log(log(N)) = 2.38

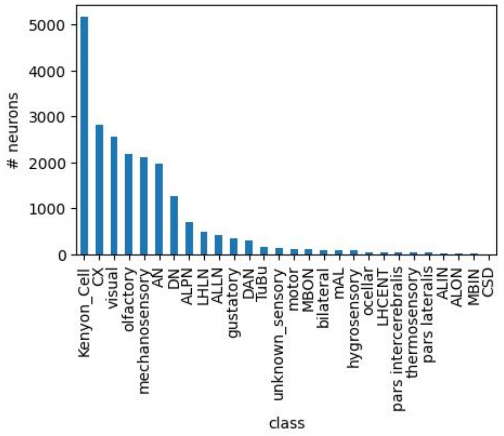


Node classification

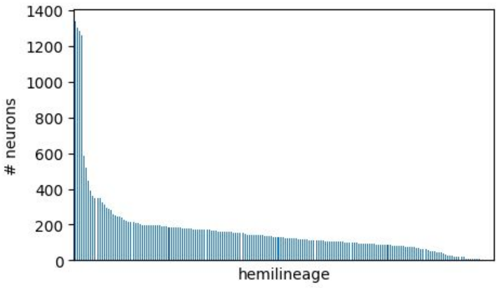
Neuron superclasses



Neuron classes



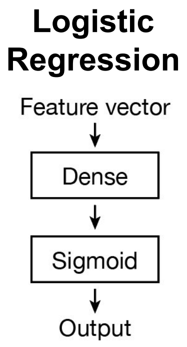
Hemilineages

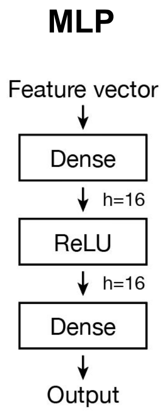


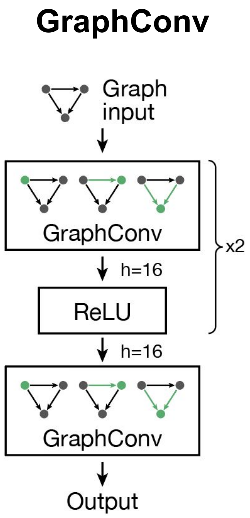
Tasks: predicting (the major classes of):

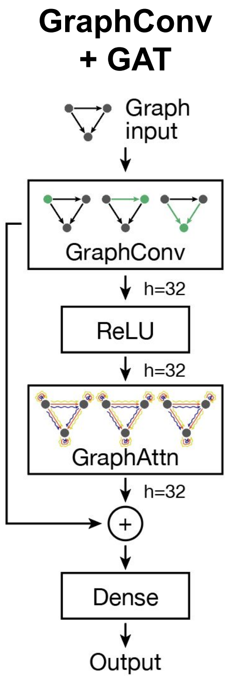
- **Neuron superclass:** coarse classification of neurons
- **Neuron class:** finer classification of neurons
- **Hemilineages:** neurons from the same hemilineage came from the same stem cell

Model architecture









We compared:

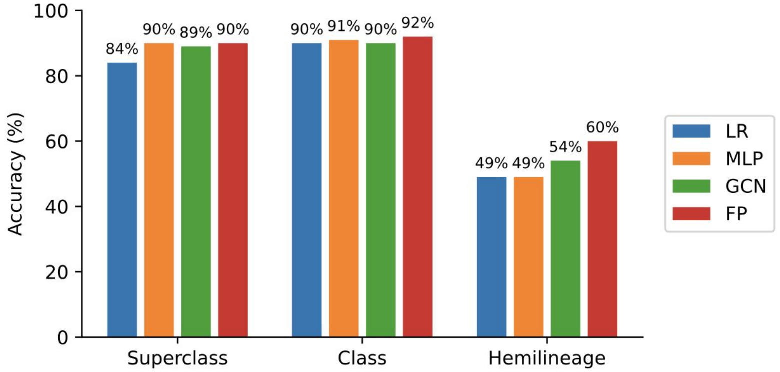
- 2 non-GNN baselines:
  - Linear regression (LR)
  - 3-layer perceptron (MLP)
- 2 GNN models:
  - Graph Convolutional Network (GCN)
  - “Fast-Prepared” (FP): GraphConv + Graph Attention layer (Deac 2019)

Results

best

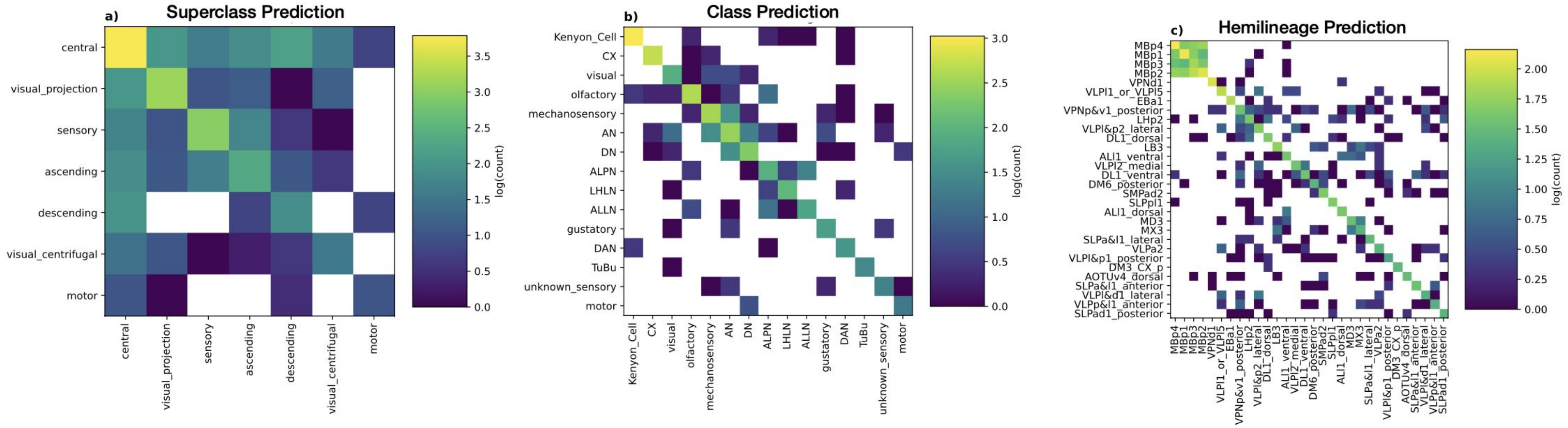
Task	Features	LR	MLP	GCN	FP
Superclass prediction	Node2Vec	82% (0.51)	89% (0.60)	89% (0.55)	**90% (0.66)**
	Spectral	76% (0.34)	86% (0.49)	**89%** (0.58)	89% **(0.62)**
	Both	84% (0.56)	90% (0.59)	89% (0.62)	**90% (0.65)**
Class prediction	Node2Vec	89% (0.81)	91% (0.82)	90% (0.83)	**92% (0.86)**
	Spectral	73% (0.53)	89% (0.80)	**92% (0.87)**	91% (0.84)
	Both	90% (0.84)	91% (0.84)	90% (0.84)	**92% (0.87)**
Hemilineage prediction	Node2Vec	47% (0.53)	48% (0.55)	53% (0.57)	**57% (0.63)**
	Spectral	30% (0.27)	41% (0.41)	52% **(0.57)**	**53%** (0.56)
	Both	49% (0.56)	49% (0.57)	54% (0.58)	**60% (0.66)**

(using spectral+Node2Vec features)



- All models (even linear baseline) performs well!
- Hemilineage prediction (30-class) is the hardest
- Conv+GAT (FP) model performs the best
- Spectral features don’t add major information on top of Node2Vec features
- **More complex models have a bigger advantage for the harder tasks**

Results



Confusion matrix (log scale) for all classification tasks

Observation: some classes are more easily confused

Discussion

- Conclusions:
    - The *Drosophila* central brain is roughly a (directed) scale-free network
    - Node embedding based on connectivity (spectral/node2vec) enables prediction of neuron features based on neighborhood information
    - Complex GNN models (GraphConv+GAT) effective at predicting neuron attributes
    - The advantage of more complex models is more pronounced in harder tasks
    - The methodology and approach can be extended to other organisms and brain regions.
- Limitations & future work:
    - Incomplete labeling of neuron classifications and hemilineages in the dataset.
    - Improving proofreading for optic neurons